DOCUMENT RESUME

ED 174 990 CS 204 988

AUTHOR McElreath, Mark P.
TITLE New Tools for Precision Journalists: Using Stems and Leaves, Boxes and Whiskers--and a Few Flogs--to Describe Recent Trends in Journalism Graduates.
PUB DATE Aug 79
NOTE 26p.: Paper presented at the Annual Meeting of the Association for Education in Journalism (62nd, Houston, Texas, August 5-8, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; *Graduate Surveys; Graphs; *Journalism; *Nonparametric Statistics; *Statistical Analysis; *Statistical Studies
IDENTIFIERS *Precision Journalism

ABSTRACT

Graphic statistics as developed by John Tukey have some practical benefits for precision journalists as they assist in spotting trends not as easily noted without the use of nonparametric statistics. Some especially helpful graphic statistics are: "stem-and-leaf display," "box plots," "froots," and "flogs." These can be used to see relationships in statistics related to data from surveys of journalism graduates, including information about salary ranges, job search time, time on the job market as related to year of graduation or to sex, and number of journalism graduates with specific majors. (Fifteen tables are included to show how the statistical techniques were applied to the survey data.) (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED174990

NEW TOOLS FOR PRECISION JOURNALISTS: USING STEMS AND LEAVES, BOXES AND WHISKERS--AND A FEW FLOGS--TO DESCRIBE RECENT TRENDS IN JOURNALISM GRADUATES

Mark P. McElreath

Assistant Professor

College of Journalism

University of Maryland

College Park, Maryland 20742

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mark P. McElreath

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented to the Theory and Methodology Division, Association for Education in Journalism Convention, Houston, Texas, August 1979.

CS204 988

NEW TOOLS FOR PRECISION JOURNALISTS: USING STEMS AND LEAVES, BOXES AND WHISKERS--AND A FEW FLOGS--TO DESCRIBE RECENT TRENDS IN JOURNALISM GRADUATES

There has been renewed interest lately in graphic statistics. Going beyond the standard histogram and pie chart, there have been efforts to generate multivariate plots,[1] cartoon faces to represent multivariate data,[2] and coded matrices to represent multiple patterns, particularly the new mapping procedures used by the U.S. Bureau of Census.[3] Some of these new graphic statistics, developed by John Tukey and others, may be appropriate not only for social scientists, they may have practical benefits for precision journalists.[4] The Tukey technique requires graph paper and pencils. Computer programs have been developed to yield the Tukey nonparametric statistics; however, the Tukey graphic statistics may be appreciated most by those not wanting to bother with a computer but who want to look at a bunch of numbers and within a short time "see" trends and relationships.

Some of the Tukey nonparametric statistics will be used in this paper to explore three data sets of interest to journalists--recent nationwide surveys of journalism graduates conducted by The Newspaper Fund, specifically the 1975, 1976 and 1977 surveys.[5] Each year a stratified, clustered and random sample of journalism colleges and universities in the United States are contacted, and a proportionate sample of their recent graduates are mailed a questionaire.

Because fairly large samples result, and because the variance in previous survey results can be analyzed, it would be possible in certain instances to use parametric statistics in our analysis. However, the purpose of this paper is to present new nonparametric graphic statistics that precision journalists and communication researchers may want to use in their own work. As any nonparametric statistical procedure, the Tukey technique for exploratory data analysis can be used to identify trends and possible relationships without making any assumptions about the distribution on any one measure.

Tukey has developed new graphic statistics--and also a new way to tally a bunch of numbers. One fast way of "scratching down numbers" is to use one pencil stroke per item, counting in clusters of five (𝍸). Working quickly with this tallying procedure, you can make mistakes (𝍸 or 𝍸𝍸, for example). Tukey suggests a simple method of dots and connecting lines, counting in clusters of ten. One is . Four is ⁘ Eight is ⊡ Ten is ⊠ The order in which the four dots are placed makes no difference, nor does it matter in which order you connect the dots. With practice, fewer mistakes are made tallying with the Tukey technique. Table I shows how the technique was applied to the salary ranges of the 1977 journalism graduates.

Table I also shows what Tukey means by a stem-and-leaf display.[6] The stem is the vertical column representing the basic units of the numbers being examined. The leaves are the final digits or actual counts in the various catagories in the stem. Tallying a bunch of

numbers in a stem-and-leaf format allows you to quickly "get a feel" for what lies within the data. Some have referred to Tukey's stem-and-leaf displays as "hanging" histograms;[7] and rightly so, for the end result is a bar graph of the frequencies of counts (the leaves) within categories (in the stem). There is an important advantage to the stem-and-leaf technique, especially when you are initially "just looking at" some numbers--the histogram pattern emerges as you are scratching down numbers.

Table 2 shows back-to-back stem-and-leaf displays of salary ranges for men and women in 1975. Apparently there is no major difference between the two sets of data; and confirming this fact, a chi square test indicates no significant interaction between sex and salary for the 1975 graduates.[8]

Table 3 shows how the stem can represent precise units. The table shows the size of 1977 graduating classes at all journalism schools in the United States. The stem contains values ranging from 0 to 640. Notice how the one's digit is contained in the leaf display and the ten's digit and 100's digit is placed within the stem. The names of the larger schools are included in the table to give even more meaning to the data. It is apparent that there are quite a number of small graduating classes and just a few big ones.

Because what seems to be interesting in this bunch of numbers is at extremes, there is a better way to display the data. Table 4 contains the same data as in Table 3, only the data are reexpressed as common logarithms. Now the extreme values "pop out" of the picture. The names of the smaller and bigger schools make the

picture even more meaningful.

Using the letter codes for frequency counts in stem-and-leaf displays helps sometimes. Table 5 shows the amount of time 1976 journalism graduates took to find a media job, with letter codes for each major sequence. The fairly equal distribution of majors "in the leaves" indicates a graduate's specific major may not be an important factor in the amount of time it takes him or her to find a media job--it might be wise to look for other possible reasons, such as the student's grade point average, social economic status, or whether the student has an internship or not. Table 5 suggests that the student's major is not that important in determining how long it might take for the graduate to find work.

Stem-and-leaf displays can be transformed into what Tukey calls "box plots." A box plot represents median and quartile values within the data set. The upper quartile is a point below which 75% of the data fell; and the lower quartile is a point below which 25% of the data fall. The median is shown as a horizontal bar across the box. A box plot graphically shows the central tendency and dispersion of the data set. Table 6 is a box plot of entry-level salaries paid 1977 journalism graduates. Tukey calls the quartile points "hinges" because they are halfway from the middle to the ends and would serve as hinges if the stem were folded at that point. Numbers not too far "outside" the hinges are connected to the box by lines, which Tukey thinks look like whiskers.[9]

Confidence intervals around the median can be approximated on a box with notches. Table 7 is a notched box plot of salary levels for 1975, 1976 and 1977 journalism graduates.[10] The overlapping notches

for 1975 and 1976 box plots indicate no major shift in the median salary level during that two-year period; however, the size of the upper quartile for the 1976 data set indicates an upward turn in salary levels. In 1977 this upward trend continued with the median salaries going from approximately $155 per week to approximately $185 per week.

The differences in actual numbers within each data set can be shown in box plots by making the sizes of the boxes in proportion to the square root of the total number or sample size. Other expressions of the differences in sample size are possible (for example, making widths of boxes directly proportional to size); yet, as Tukey and his associates point out, many variability measures are proportional to the root of group size (for example, standard error) and widths directly proportional to size often overemphasie differences.[11] Table 8 is a variable width notched box plot showing salary levels for white and minority journalism graduates in 1977. The overlapping notch points would indicate most minorities and whites received comparable entry-level salaries, all other factors being equal (which, of course, they never are). The skinniness of the minority box warns you to be cautious in drawing conclusions.

Simple fractions are the workhorses of the precision journalist--they are easy to calculate and easily understood. Tukey has developed a procedure that greatly increases the usefulness of simple fractions for data analysis. He transforms them into pluralities, metaphorically folding one side of the fraction onto the other side and looking at the difference. If f is the fraction of interest, a plurality

equals f minus (1-f). In this way, Tukey creates a scale with a value of zero when f=1/2. A "folded fraction" using percentages -- for example, (% yes) minus (% no) -- Tukey calls an "fcent." If 50% is the fraction of interest, the fcent value is zero. If more than 50%, then the fcent has a positive sign; if less than 50%, the fcent has a negative sign.

Tables 9, 10 and 11 show the actual counts and pluralities of the amount of time men and women journalism graduates spent on the job market looking for a media job in the years 1975, 1976 and 1977. When working with percentages below 20% or above 80%, Tukey recommends reexpressing the data in terms of logarithms or square roots; for such reexpressions "stretch out the ends" of the scale, making extreme values more readily apparent. (Refer again to table 3 and 4 to see the effect of logarithm reexpressions of extreme values.) Tukey has developed formulae for folding fractions re-expressed as square roots and logs, noting that they could be called "folded roots" and "folded logs"--but, in his own inimitable style, he calls them "froots" and "flogs." Tables 9, 10 and 11 also show the basic data reexpressed as froots and flogs,

Table 12 combines the best of Tables 9-11, showing the amount of time men and women journalism graduates spent on the job market, expressed in flogs, for 1975, 1976 and 1977. Note that a zero flog means no difference in men and women job search time; positive flogs indicate more men than women in the particular category; negative flogs indicate more women than men in the particular category of search time. The floggram (Table 12) indicates a very slight

tendency for there to be more women taking longer to find a media job than men; and conversely, more men finding media jobs quicker than women. However, the tendency is very slight. By looking at the symbols used for each year's data points, trends for each year can be seen. In 1976 there was a definite trend for women to be on the job market longer than men; but, for the other years, the trend is not so obvious. Confirming this, chi square tests for independence of sex and job search time for 1975 and 1977 journalism graduates were not significant; while the chi square test for 1976 graduates indicated the variables were not independent.[13]

To keep matters in perspective, it is important to note that floggrams sometimes do not greatly improve upon standard bar graphs; and Tukey has acknowledged this stating, "Sometimes going to flogs makes things wonderful; sometimes merely good, sometimes no better, sometimes worse."[13] Tables 13 and 14 should be viewed together. Table 13 shows the percentage of graduates with specific majors in journalism for the years 1975-1977. Table 14 shows the same data reexpressed as flogs. The slight increase in 1976 in the numbers of graduates majoring in magazine journalism and photojournalism "pops out" more in the floggram; but, not much more can be learned from the floggram than from the percentage bar graph. It is apparent that there are fewer broadcasting graduates, more graduates with "General" majors, and a definite increase in the numbers of news editorial graduates. The floggram, Table 14, also shows this. Flogs and percentages, in this case, are equally as useful for highlighting trends in the data. In his own defense, Tukey says: "On balance, though, (flogs) seem to help often

enough so that the writer usually tries flogs even before percents or pluralities--flogs are the first-aid bandages for counted fractions."[15]

The last table, Table 15, is a floggram showing the difference it makes in the amount of time journalism graduates spend on the job market if they have or have not had an internship as an undergraduate. It is readily apparent that having an internship as an undergraduate greatly increases the likelihood that the student will find a media job quickly; and without an internship, it definitely takes longer to find a media job.[16] This is a good example of when one of Tukey's folded fractions is a very good way to display data.

It has not been the purpose of this paper to integrate these findings with previous research or with theories regarding journalism education and newswork in general. Rather, the purpose has been to explore possible trends and relationships and to explain how some of the new Tukey nonparametric statistics can be used to analyze and display trends and relationships in data sets of interest to journalists.[17]

Table 1. Tukey tallying: Salary ranges of the 1977 journalism graduates

Less than $100	
$101 – $110	1:
120	1:
130	𝍸
140	𝍸 ⊏
$141 – $150	𝍸 𝍸.
160	𝍸 𝍸::
170	𝍸 𝍸 ⊏
180	𝍸 𝍸 1:
190	𝍸 𝍸 𝍸
$199 – $200	𝍸 ⊏
210	𝍸 ⊏.
220	𝍸 𝍸
230	𝍸 :·
240	𝍸
250	1:
More than $250	

Table 2. Back-to-back stem-and-leaf displays of male and female salary levels for 1975 journalism graduates

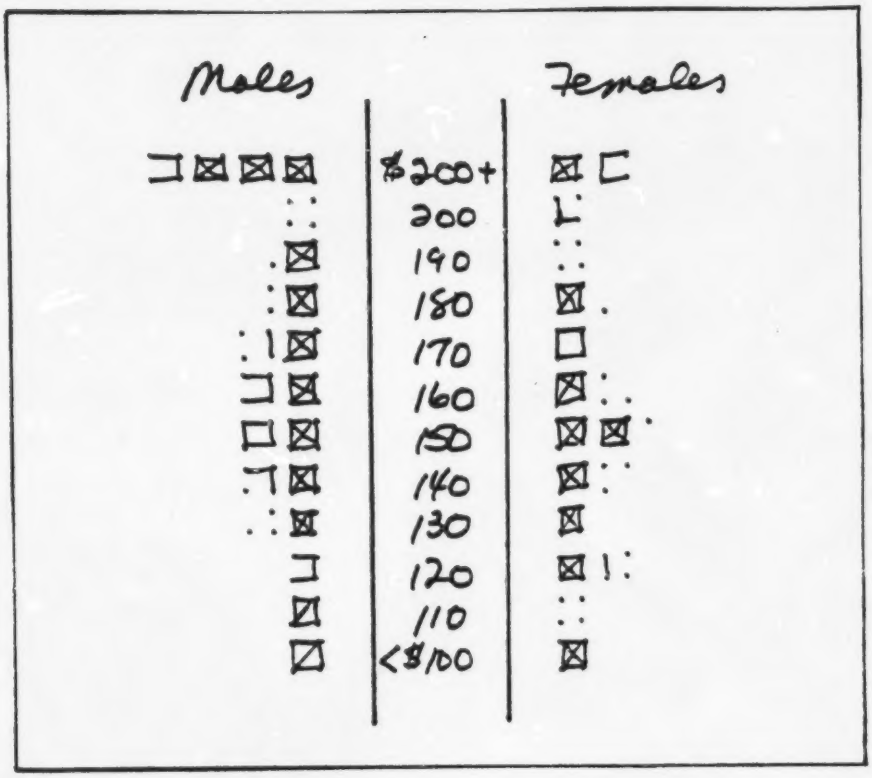


Table 3. Size of journalism graduating classes in 1977

0	9965618322392583774517 52	
1 *	722457764264806 14	
2	07801305950876351697277 30	
3	3689462670	
4	07754728067320 15	
5 *	2902587766 29	
6	60120211	
7	315149938	
8	693	
9	6712	
10 *	85052	
11	27863348	
12	8750	
13	424192	
14	5261	
15 *	821	
16	7	
17	25	
18	4	
19		
20 *	48	Calif. State, Fullerton; Washington
21	3	South Carolina
22	2	Michigan State
23		
24	2	Alabama
25 *		
26		
27	0	Kansas
28		
29		
30 *		
31		
32	9	Missouri
33		
34		
35 *	859	Boston U.; Georgia; Syracuse
47 *	9	Florida
64 *	7	Texas @ Austin

Table 4. Stem and leaf display of log of the size of journalism graduating classes in 1977

Size		
0.0	00	(N=1) Culver-Stockton (Mo); Texas Eastern
0.1		
.2		
.3	0000	(N=2) Wyoming; MacMurray; Henderson St.; Hastings
.4	777	(N=3) Okla. Baptist; Keene St (N.H.); Goshen College
.5		
.6	06.9	
.7	800	
.8	444	
.9	55050	
1.0	77704	
.1	47444	
.2	3330060	
.3	00209906092460	
.4	3464311633337	
.5	15893505 6	
.6	07754728067320159	} range = 40-62
.7	17146554417789888	
.8	16569969	
.9	3718856	
2.0	32021467655 6798	} range = 100-155
.1	00223242656598 7	
.2	2346	
.3	11248	
.4	3	(N=270) Kansas
.5	1555	Missouri; Georgia; Boston U.; Syracuse
.6	8	(N=479) Florida
.7		
.8	1	(N=647) Texas @ Austin

Table 5. Search time for a media job by recent graduates with a specific major, 1976

Search time	major code*
Before graduation	NNNNNNNNNNnBbAaPRm
1-30 days	NNNnbaprmfo
31-60 days	NNnbaprmfgo
61-90 days	NnbAamgo
91 days to 6 months	NNnbaprmfg
6-12 months	nbaprmf
More than 1 year	nfo

*Major code: N=News Editorial
B=Broadcasting
A=Advertising
PR=Public Relations
M=Magazines
F=Photojournalism
G=General
O=Other

* Note: capital letters represent units of ten; lower case letters, less than ten.

Table 6. Box plot of entry-level salaries paid 1977 journalism graduates

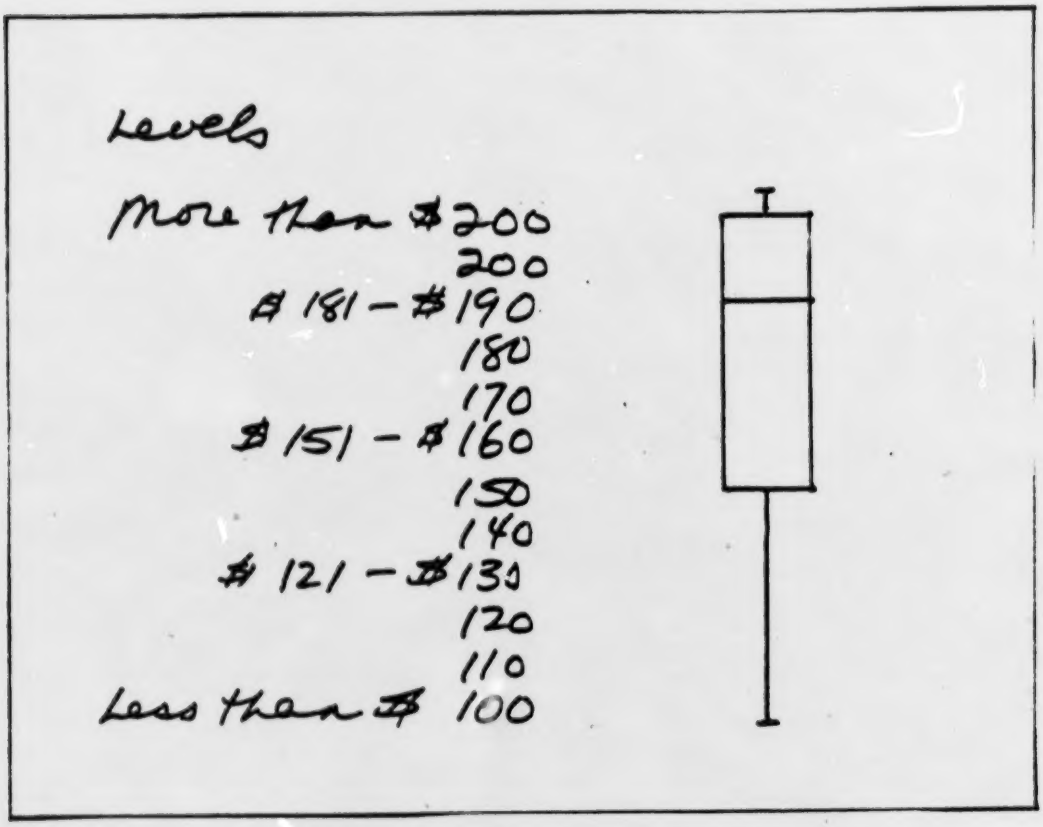


Table 7. Notched box plot of salary levels for 1975, 1976, 1977 journalism graduates.

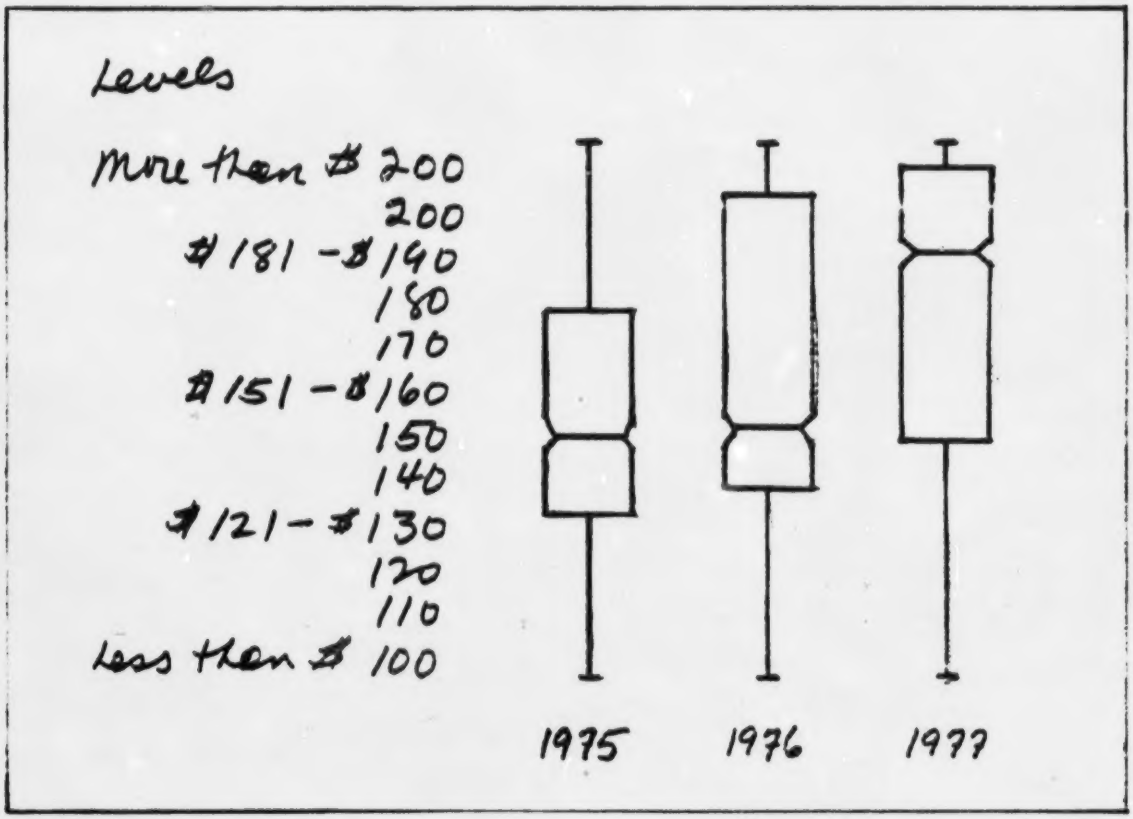


Table 8. Variable width notched box plot of salary levels for white and minority journalism graduates in 1977

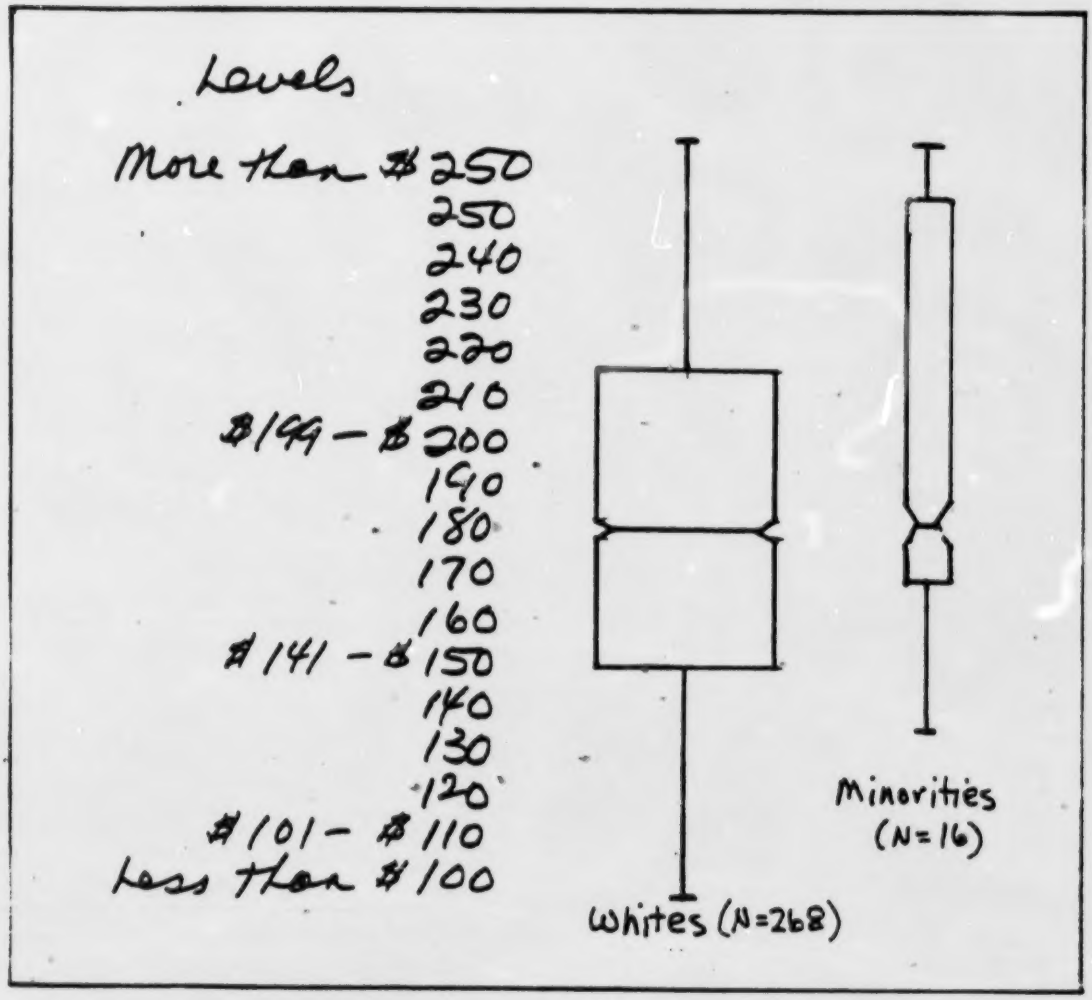


Table 9. Time on the job market for men and women journalism graduates in 1977

Search time to find a job	Actual Counts Women	Men	froots	flogs
Before graduation	60	67	-.06	-.05
Less than 1 month	41	25	.24	.24
1-3 months	22	23	-.02	-.02
3-6 months	24	11	.38	.38
6-12 months	16	17	-.03	-.03
More than 1 yr.	4	7	-.27	-.27
	N=167	150		

Table 10. Time on the job market for men and women journalism graduates in 1976

Search time to find a job	Actual Counts Men	Actual Counts Women	flogs
Before graduation	98	60	.24
Less than 1 month	28	33	-.08
1-3 months	39	48	-.11
3-6 months	17	27	-.23
6-12 months	4	11	-.50
More than 1 year	2	2	.00

Table 11. Time on the job market for men and women journalism graduates in 1975

Search time to find a job	Actual Counts Men	Actual Counts Women	flogs
Before graduation	69	41	.26
Less than 1 month	32	31	.02
1-3 months	40	41	-.01
3-6 months	26	21	.10
6-12 months	15	14	.03
More than 1 year	0	0	.00*

* An abnormal value of -1.8 occurs with the log of zero here; it has been replaced with zero. Tukey suggests adding a constant to all values to avoid such abnormalities.

Table 12. Time on the job market for men and women journalism graduates: 1975 to 1977

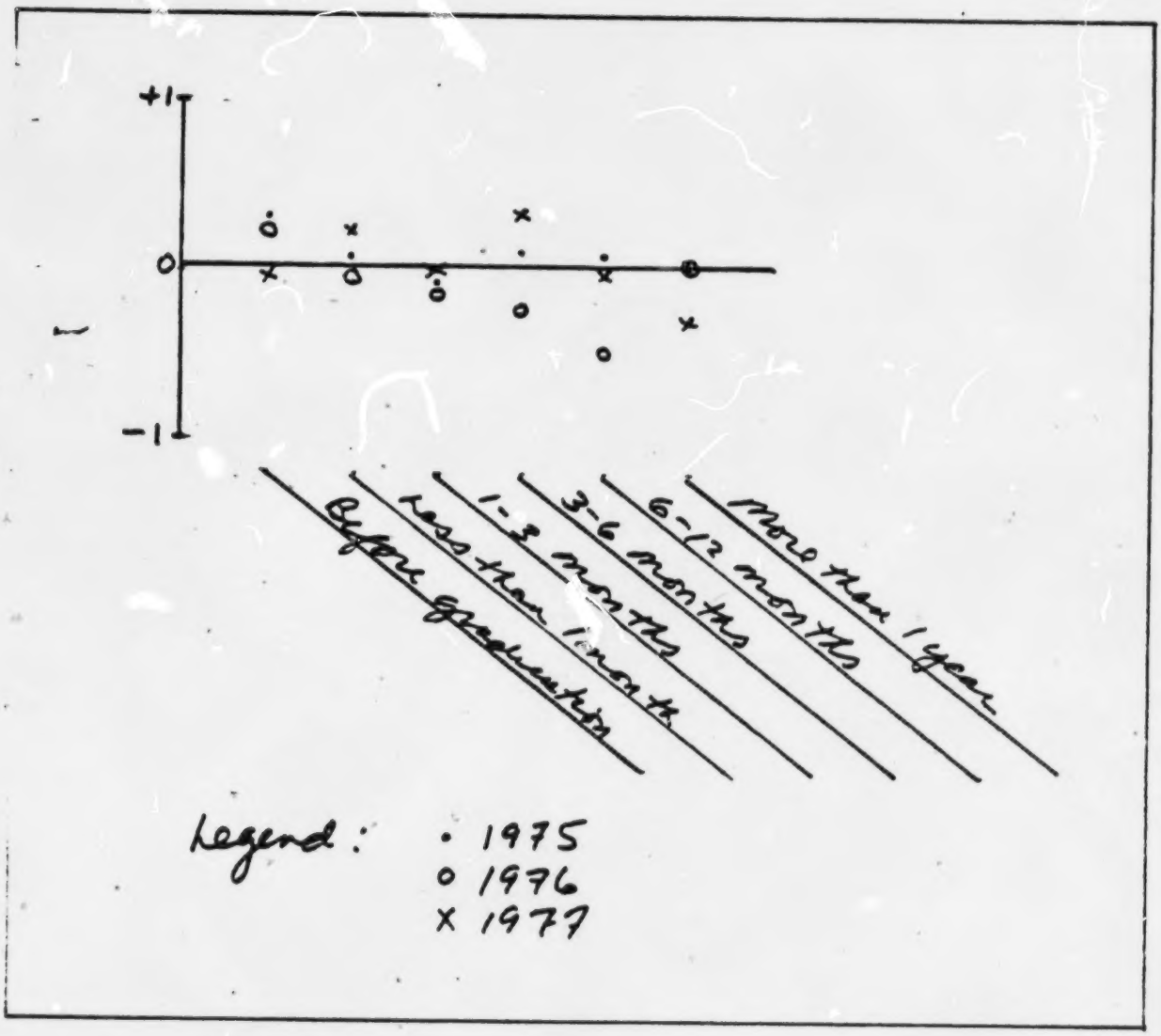


Table 13. Journalism graduates by specific majors--1975 tc 1977; expressed as percentages

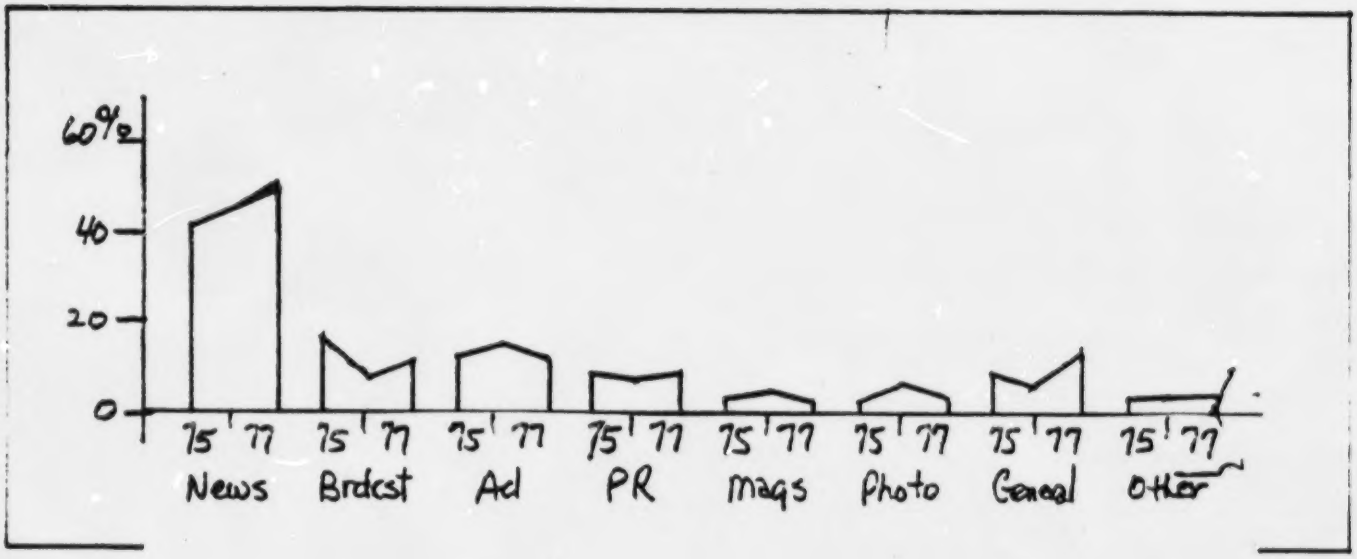


Table 14. Same data as above; expressed as flogs

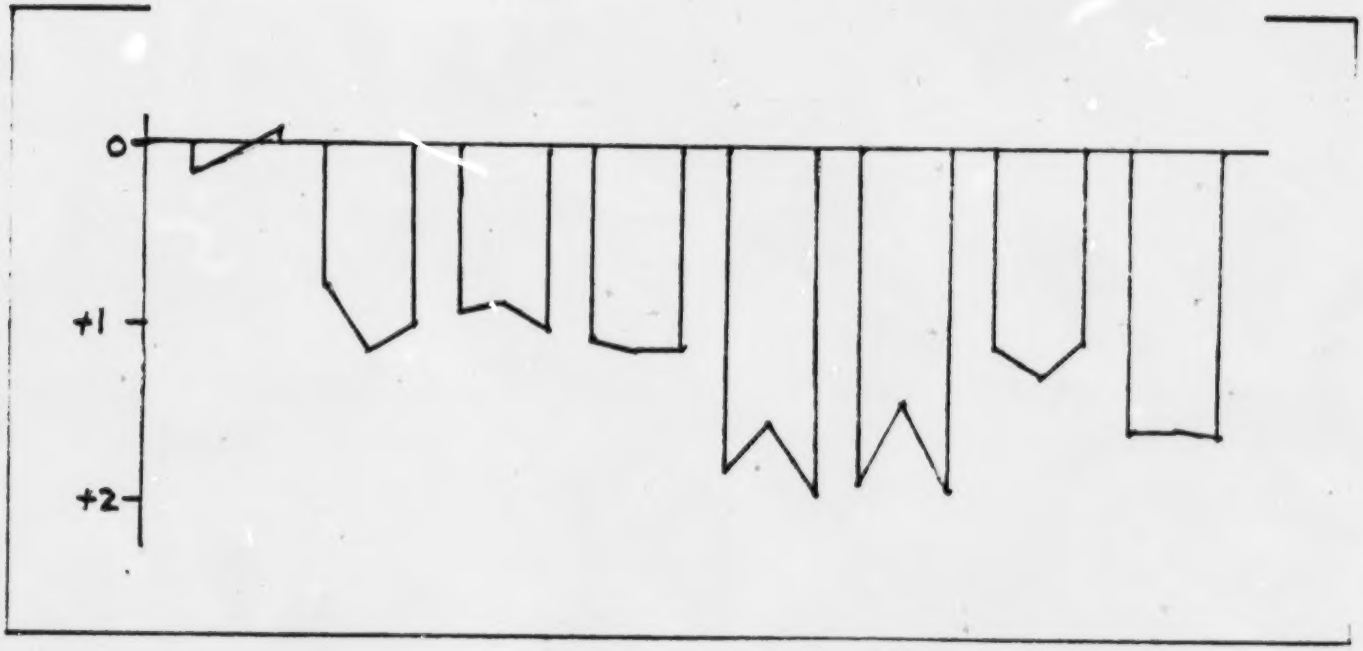
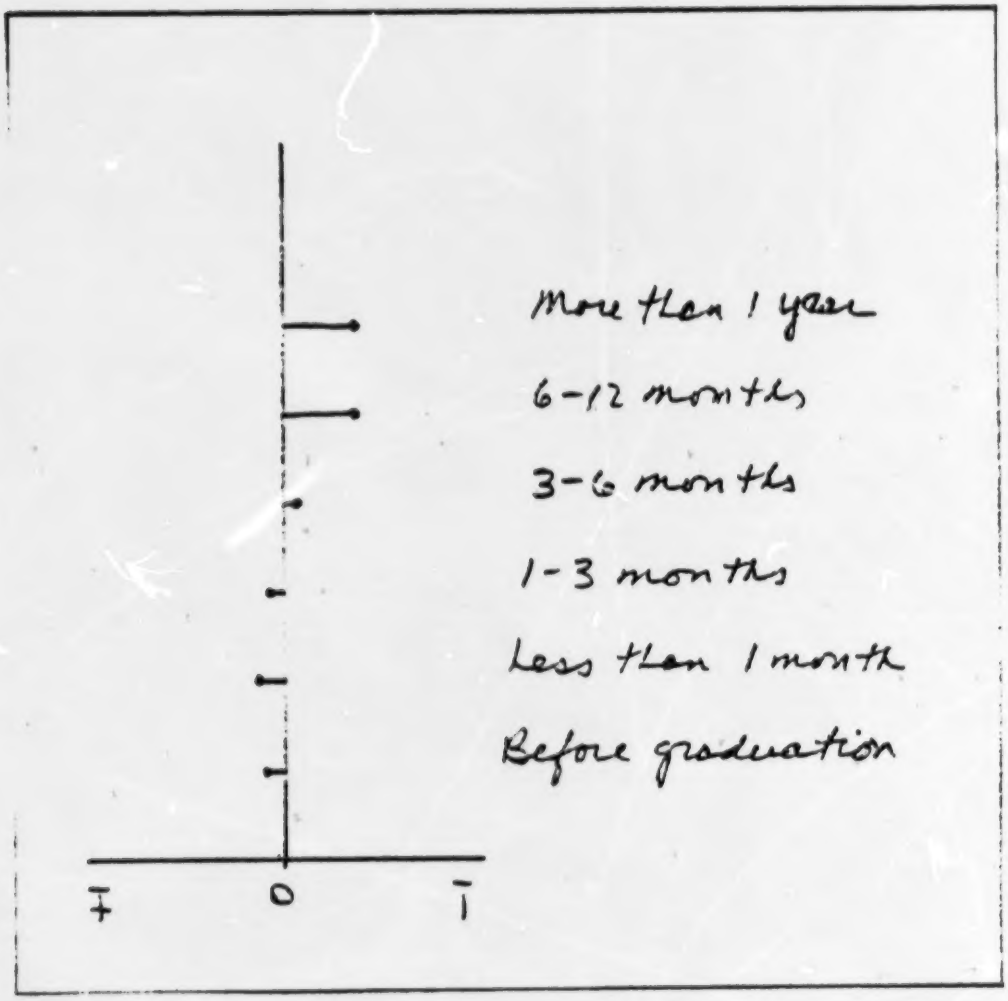


Table 15. Time on the job market for 1976 journalism graduates with and without internships prior to graduation; expressed as flogs



Footnotes

1. See D.F. Andrew, "Plots of high dimensional data," Biometrics 28 pp. 125-36, 1972, which is cited in an article by James R. Beniger, "Quantitative Graphics in Statistics: a brief history," American Statistitian 32:1 pp. 1-11 Feb. 1978. For traditional graphic statistics, see Mary Eleanor Spear, Practical Charting Techniques (New York: McGraw-Hill, 1969).
2. See Herman Chernoff, "The use of faces to represent points in K-dimensional space graphically," Journal of the American Statistical Association 68:342 pp. 361-68 June, 1973.
3. See Roberto Bachi, Graphic rational patterns: a new approach to graphical presentation of statistics (Jerusaleum: Israel University Press, 1968); and, the U.S. Bureau of Census and Manpower Administration, Urban Atlas, Series GE-80, Washington, D.C., U.S. Government Printing Office 1974.
4. Prepublication drafts of Tukey's book were hot property in academia in the past few years, assuring the significance of his book even before publication. Exploratory Data Analysis (Reading Mass: Addison-Wesley, 1977). Tukey is with Princeton University and the Bell Telephone Laboratories.
5. Copies of the annual survey and booklets based on the survey results are available from The Newspaper Fund, P.O. Box 300, Princeton, N.J. 08540. Thomas E. Engleman is Executive Director of the Newspaper Fund.
6. Tukey has a glossary in Exploratory Data Analysis, op.cit., with dozens of special and unique terms that he uses in a refreshing way to describe his statistics.

7. Beniger, op.cit.

8. Chi square = 15.50, with 11 degrees of freedom; significance = .16.

9. If d equals the upper quartile minus the lower quartile, you draw the whisker by connecting all data points 1.5d from the nearest edge of the box with a straight line. Points beyond 1.5d are dotted separately, clearly showing "far out" values.

10. One approximation of the standard deviation of the median was developed by Kendall and Stuart (THE ADVANCED THEORY OF STATISTICS, Vol. 1, 2nd ed., Ch. 14, New York: Hafner Publ., 1967). With R the interquartile range and N the number of observations, the standard deviation (s.d.) is given by the formula: $s.d. = 1.25R/1.35\sqrt{N}$. To calculate notches indicating a 95 percent confidence interval about each median, you should multiply the standard deviation by 1.96. Tukey and his associate (see footnote 11) point out that 1.96 is appropriate only if the s.d. of the two groups are vastly different, and they suggest a lower figure of 1.7 as being more appropriate. (The 1.96 constant, Tukey says, creates a confidence interval far in excess of 99 percent.)

11. Robert McGill, John W. Tukey and Wayne A. Larsen, "Variations of Box Plots," The American Statistician, 32:1 pp. 12-16, February 1978.

12. See chapter 15 of Exploratory Data Analysis, op.cit., for froots and flogs. The formulae for froots and flogs use for the fraction of interest:

froot $= \sqrt{2f} - \sqrt{2(1-f)}$ flog $= 1/2 \log_e f - \log_e(1-f)$

13. For the 1976 data set, chi square = 17.07, with 6 degrees of freedom; significance = .009.

14. Exploratory Data Analysis, pp. 509.

15. Ibid.

16. Chi square = 13.19, with 6 degrees of freedom: significance = .04.

17. Not discussed in this paper are other Tukey nonparametic statistics, such as product ratio plots, three and more-way analyses, curve fitting and smoothing. The interested reader is referred to Exploratory Data Analysis, op.cit. See also, Data Analysis and Regression: A second course in statistics, by Frederick Morteller and John Tukey (Reading, Mass.: Addison-Westley, 1977); and, Interactive Data Analysis, A practical primer, by Donald R. McNeil (New York: John Willy & Sons, 1977), which contains special Fortran and APC computer programs for the Tukey nonparametrics.